## REMARKS

Claims 1-25 are pending in the application. Claims 1-25 have been rejected.

The title of the invention was objected to for not being descriptive. Applicants disagree because the title is descriptive of the invention claimed. Independent claims 1 and 14 recite a computer method and apparatus for collecting (e.g., searching for and extracting) people and organization information from web sites. The foregoing claim Amendments further make this clear. Thus, the present title indicates the invention to which the claims are directed as prescribed in MPEP §606.01.

The abstract was also objected to for using the term "means". The abstract has been amended to remove the term "means." The abstract has further been amended to include the language "and apparatus" and to a correct grammatical error by deleting the apostrophe in the term "URL's".

The corrections requested in paragraphs 4 and 5 of page 2 of the office action at hand are being made in the specification as set forth above. Applicants thank the Examiner for pointing out these errors. No new matter is introduced by way of these amendments.

Claims 1-25 have been rejected under 35 U.S.C. §102(b) as being anticipated by Sahami et al. "SONIA: A Service for Organizing Networked Information Autonomously" (pages 200-209).

The present invention relates to a method for collecting (e.g., searching for and extracting) people and organization information from web sites. According to this method a web site having multiple web pages is accessed. If it is determined that the web site is not a duplicate of a previously processed web site, then the method determines a subset of the multiple web pages associated with the web site to process. For each web page in the subset, the type of content found in the web page is determined. Finally, based on the determined content type, people and organization information is extracted from the web page. In this way, the number of web pages to process decreases substantially. Moreover, by knowing the type of content found on the web page, efficiency can be achieved because different types of web sites require different frequencies of crawling and different crawling and extraction strategies.

In contrast, Sahami et al. disclose a system for dynamically categorizing documents based on the text of articles that are retrieved in response to a user's query. Referring to Fig. 2 of Sahami et al., which shows the processing stages in the system referred to as "SONIA", the full text of documents are retrieved by a parallelized crawler module. The retrieved document texts are parsed, and initially, two feature selector modules prune the parsed words. The first feature selector removes all stop words and the second feature selector applies Zipf's law, which eliminates terms that appear frequently or infrequently in the entire collection of documents. If an existing profile from a previous classification scheme is used, a third feature selector module based on information theory finds terms that are most discriminating between groups of documents in any given profile. Then, the classifier categorizes documents using the existing profile and using a classification algorithm based on bayesian networks. Finally, a descriptor extractor module extracts descriptors from the document subsets so that the user can be presented with topics corresponding to the document subsets.

If the user chooses not to use an existing profile, then a third feature selector module based on entropy is used. The documents are then classified using a clustering technique. The resulting document subsets are applied to the descriptor extractor module to extract descriptors for presentation to a user.

Unlike Applicants' system, which accesses a web site having a plurality of web pages, Sahami et al. disclose a system which receives for processing a list of document identifiers such as URLs for web pages (see page 202). Thus, Sahami et al. do not disclose the claimed "accessing a web site of potential interests, the web site having a plurality of web pages" set forth in base Claims 1 and 14 of Applicants' disclosed invention.

Moreover, unlike Applicants' system, which determines a subset of multiple web pages in a web site to process, Sahami et al. utilize a clustering technique to create a novel topical categorization of all documents given as input to the SONIA system (see page 204). Also, the descriptor extractor module of the SONIA system extracts descriptors from all the document subsets output from the Clusterer module (see page 205). Thus, Sahami et al. do not anticipate the claimed "determining a subset of the plurality of web pages to process" set forth in base Claims 1 and 14 of Applicants' disclosed invention.

Therefore, the present invention as claimed in base claims 1 and 14 is not believed to be anticipated by Sahami et al. Claims 2-13 are dependent on claim 1 and claims 15-25 depend from claim 14. Thus, for at least the same reasons, dependant claims 1-13 and 15-25 should be allowed over the cited prior art.

According to the foregoing, Applicants respectfully request that the rejection under 35 U.S.C. § 102(b) in view of Sahami et al. should be withdrawn.

## CONCLUSION

In view of the above amendments and remarks, it is believed that all pending claims (claims 1-25) are in condition for allowance, and it is respectfully requested that the application be passed to issue. If the Examiner feels that a telephone conference would expedite prosecution of this case, the Examiner is invited to call the undersigned.

Respectfully submitted,

HAMILTON, BROOK, SMITH & REYNOLDS, P.C.

By 

Mary Lou Wakimura
Registration No. 31,804
Telephone: (978) 341-0036
Facsimile: (978) 341-0136

Concord, MA 01742-9133
Dated: 11/1/04